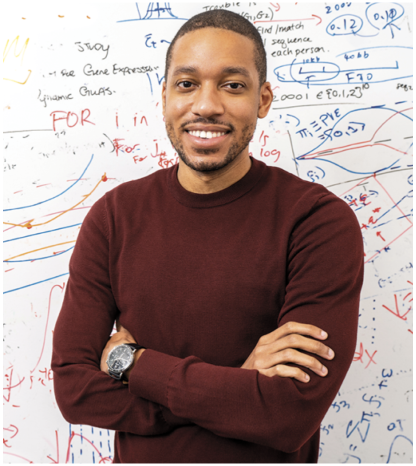
Randomness and Statistical Inference of Shapes via the Smooth Euler Characteristic Transform

Kun Meng
(Mike)

Division of Applied Mathematics
Brown University

June 27, 2023

joint work with them



Prof. Ani Eloyan
(Biostatistics@Brown)

Prof. Lorin Crawford
(Microsoft & Biostatistics@Brown)

Mr. Jinyu Wang
(Data Science@Brown)

1. Motivation (from the morphology viewpoint)

2. Statistical Methodology (not mathematically rigorous)

3. Applications

4. Mathematical Foundations

Section 1: Motivation.

(from the morphology viewpoint)

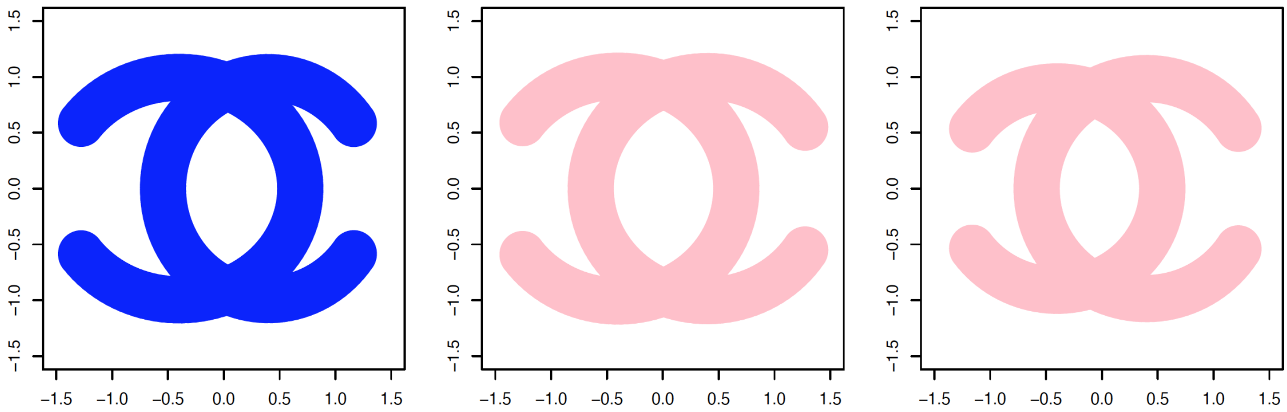
Motivation: Are the teeth from the same species?

Figure 1: We have four collections of **teeth** $\Longrightarrow$ four groups of **shapes**.

Question: Are the four collections of teeth from the same species?

Toy example: Are the shapes from the same distribution?

- $\mathbb{P}^{(1)}$ and $\mathbb{P}^{(2)}$ are shape-generating distributions, i.e., they generate "shape-valued" random variables.
- 100 blue shapes $\overset{iid}{\sim} \mathbb{P}^{(1)}$; 100 pink shapes $\overset{iid}{\sim} \mathbb{P}^{(2)}$;



Question: $\mathbb{P}^{(1)} = \mathbb{P}^{(2)}$?

To be equal, or not to be, that is the hypothesis testing question.

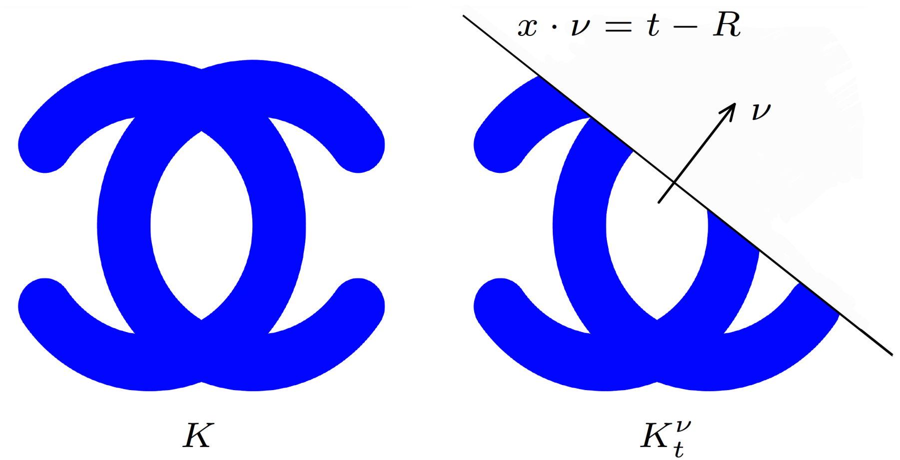
Section 2: Statistical Methodology.

(not mathematically rigorous)

Smooth Euler[1] Characteristic Transform (SECT)

- Each shape is denoted by K.
- We assume $K \subset B(0,R) = \{x \in \mathbb{R}^d : \|x\| < R\}$ which denotes a closed ball centered at the origin with a prespecified radius $R > 0$.
- For each direction $\nu \in \mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\| = 1\}$, we define a collection of sublevel sets of K by

$$K_t^\nu \overset{\text{def}}{=} \{x \in K | x \cdot \nu \leq t - R\}, \quad \text{for all } t \in [0,T], \quad \text{where } T = 2R.$$



[1] 'oi-ler'

Smooth Euler Characteristic Transform (SECT)

- We have the Euler characteristic transform (ECT)[2] of shape K

$$\begin{aligned} \mathrm{ECT}(K): \ & \mathbb{S}^{d-1} \times [0, T] \to \mathbb{Z}, \\ & (\nu,\, t) \mapsto \chi(K_t^\nu), \\ & \chi(K_t^\nu) \overset{\text{def}}{=} \text{ the Euler characteristic of } K_t^\nu. \end{aligned}$$

  E.g., Euler characteristic (a mesh) = #vertice−#edges+#faces.
- The smooth Euler characteristic transform (SECT)[3] is defined as

$$\mathrm{SECT}(K): \ \ \mathbb{S}^{d-1} \times [0, T] \to \mathbb{R}, \quad (\nu, t) \mapsto \mathrm{SECT}(K)(\nu; t),$$

$$\text{where } \ \mathrm{SECT}(K)(\nu; t) \overset{\text{def}}{=} \int_0^t \chi(K_\tau^\nu)\, d\tau - \frac{t}{T} \int_0^T \chi(K_\tau^\nu)\, d\tau.$$

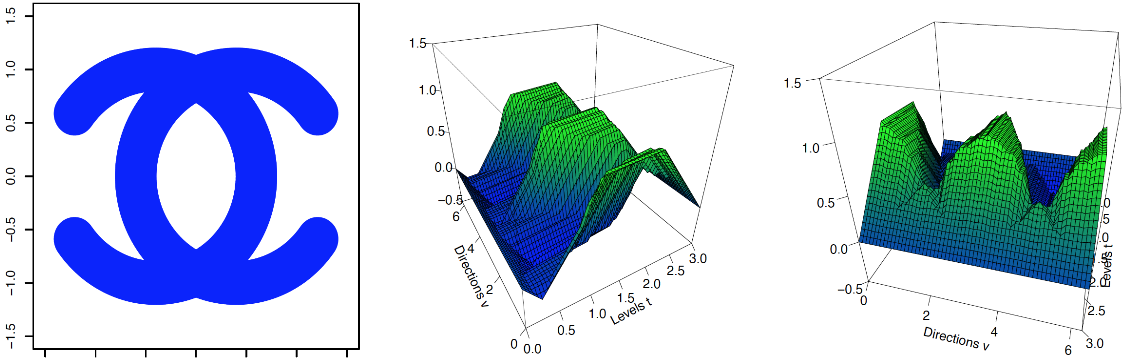
---

[2] K. Turner, S. Mukherjee, and D. M. Boyer. Persistent homology transform for modeling shapes and surfaces. *Information and Inference: A Journal of the IMA*, 3(4):310–344, 2014b.

[3] L. Crawford, A. Monod, A. X. Chen, S. Mukherjee, and R. Rabadan. Predicting clinical outcomes in glioblastoma: an application of topological and functional data analysis. *Journal of the American Statistical Association*, 115 (531):1139–1150, 2020.

Smooth Euler Characteristic Transform (SECT)

Shape $\rightarrow$ scalar field on $\mathbb{S}^{d-1} \times [0, T]$



Scalar field $(\nu, t) \mapsto \mathrm{SECT}(K)(\theta; t)$, where $\nu = (\cos\theta, \sin\theta) \in \mathbb{S}^1$.

Shape analysis $\rightarrow$ manifold learning[4]

Imaging data $\rightarrow$ functional data

[4] **Kun Meng** and Ani Eloyan. Principal manifold estimation via model complexity selection. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 83(2):369, 2021.

Smooth Euler Characteristic Transform (SECT)

- The shape-to-SECT map is invertible[5], which was proved using o-minimal structures[6] and the Schapira's inversion formula[7].
- Therefore, SECT(K) preserves all the information of the shape K.

Contribution of this project: If the shape K is random,

- $(\nu, t) \mapsto \mathrm{SECT}(K)(\nu; t)$ is a random field indexed by $\mathbb{S}^{d-1} \times [0, T]$;
- for each direction ν, the function $t \mapsto \mathrm{SECT}(K)(\nu; t)$ is a stochastic process indexed by $t \in [0, T]$.

---

[5] R. Ghrist, R. Levanger, and H. Mai. Persistent homology and Euler integral transforms. *Journal of Applied and Computational topology*, 2, pages55–60 (2018).

[6] Lou van den Dries. Tame Topology and O-minimal Structures. *London Mathematical Society Lecture Note Series*. Cambridge University Press, 1998.

[7] P. Schapira. Tomography of constructible functions. In Applied Algebra, Algebraic Algorithms and Error-Correcting Codes: 11th International Symposium, AAECC-11 Paris, France, July 17–22, 1995 Proceedings 11, pages 427–435. Springer, 1995

Hypothesis Testing: To be equal or not to be?

- Suppose we have two underlying distributions $\mathbb{P}^{(1)}$ and $\mathbb{P}^{(2)}$ generating random shapes.
- We want to test " $\mathbb{P}^{(1)} = \mathbb{P}^{(2)}$."
- Then, we have two collections of mean functions

$$\boxed{m_\nu^{(j)}(t) = \mathbb{E}^{(j)}\{\mathrm{SECT}(\cdot)(\nu;t)\} = \int \{\mathrm{SECT}(K)(\nu;t)\}\, \mathbb{P}^{(j)}(dK),}$$

  for $j \in \{1,2\}$, $t \in [0,T]$ and $\nu \in \mathbb{S}^{d-1}$.
- We test the following weaker form using the first moments

$$H_0 : m_\nu^{(1)}(t) = m_\nu^{(2)}(t) \text{ for all } (\nu,t) \in \mathbb{S}^{d-1} \times [0,T]$$
$$vs. \quad H_1 : m_\nu^{(1)}(t) \neq m_\nu^{(2)}(t) \text{ for some } (\nu,t).$$

- One may be only concerned with the discrepancy between means. Or, from another viewpoint, rejecting H_0 implies $\mathbb{P}^{(1)} \neq \mathbb{P}^{(2)}$.

Hypothesis Testing

- The following involves infinitely many directions $\nu \in \mathbb{S}^{d-1}$

$$H_0 : m_\nu^{(1)}(t) = m_\nu^{(2)}(t) \text{ for all } (\nu, t) \in \mathbb{S}^{d-1} \times [0, T]$$
$$vs. \quad H_1 : m_\nu^{(1)}(t) \neq m_\nu^{(2)}(t) \text{ for some } (\nu, t).n$$

- Considering all the directions would induce an infeasible multiple-comparison problem.
- We only investigate the following "distinguishing direction"

$$\nu^* \overset{\text{def}}{=} \arg \max_{\nu \in \mathbb{S}^{d-1}} \left\{ \sup_{t \in [0,T]} \left| m_\nu^{(1)}(t) - m_\nu^{(2)}(t) \right| \right\}$$
$$= \arg \max_{\nu \in \mathbb{S}^{d-1}} \left\| m_\nu^{(1)} - m_\nu^{(2)} \right\|_{C([0,T])}.$$

- One may rigorously show that it suffices to focus on direction ν^*.

Covariance Kernels

- We focus on one distinguishing direction ν^*.
- Associated with distribution $\mathbb{P}^{(j)}$, for $j \in \{1, 2\}$, the stochastic process $\mathrm{SECT}(K)(\nu^*, \cdot)$ has a covariance function $\kappa_{\nu^*}^{(j)}(s,t)$.
- **Assumption:**[8] $\kappa_{\nu^*}^{(1)}(s,t) = \kappa_{\nu^*}^{(2)}(s,t) \overset{\text{def}}{=} \kappa(s,t)$.
- Under some topological conditions, the following integral operator is a Hilbert-Schmidt (hence, compact) operator on $L^2(0,T)$,

$$f \mapsto \int_0^T \kappa(s,\cdot) f(s)\, ds,$$

and it has the eigenvalues $\{\lambda_l\}_{l=1}^\infty$ and orthonormal eigenfunctions $\{\phi_l\}_{l=1}^\infty$.

[8] This assumption corresponds to the null hypothesis $\mathbb{P}^{(1)} = \mathbb{P}^{(2)}$. With a permutation trick, our statistical method is robust to the violation of the assumption.

Karhunen–Loève expansion

- We have the following Karhunen–Loève expansion

$$\mathrm{SECT}(K)(\nu^*;t) = m_{\nu^*}^{(j)}(t) + \sum_{l=1}^{\infty} \sqrt{\lambda_l} \cdot Z_l(K) \cdot \phi_l(t),$$

where $Z_l(K) := \frac{1}{\sqrt{\lambda_l}} \int_0^T \left\{ \mathrm{SECT}(K)(\nu^*;t) - m_{\nu^*}^{(j)}(t) \right\} \cdot \phi_l(t)\, dt$

- The convergence of $\sum$ is in the $L_t^\infty L_K^2(dt, \mathbb{P}^{(j)}(dK))$ topology.
- Z_l is of mean 0 and variance 1, and they are mutually uncorrelated according to $\mathbb{P}^{(j)}$ across $l = 1, 2, ...$

Hypothesis Testing

- Data: shapes $\{K_i^{(1)}\}_{i=1}^n \overset{iid}{\sim} \mathbb{P}^{(1)}$ and $\{K_i^{(2)}\}_{i=1}^n \overset{iid}{\sim} \mathbb{P}^{(2)}$.
- The Karhunen–Loève expansions provide the following

$$\begin{aligned} X_i(t) &\overset{\text{def}}{=} \text{SECT}(K_i^{(1)})(\nu^*;t) - \text{SECT}(K_i^{(2)})(\nu^*;t) \\ &= \left\{m_{\nu^*}^{(1)}(t) - m_{\nu^*}^{(2)}(t)\right\} \\ &\quad + \sum_{l=1}^{\infty} \sqrt{2\lambda_l} \cdot \left(\frac{Z_l(K_i^{(1)}) - Z_l(K_i^{(2)})}{\sqrt{2}}\right) \cdot \phi_l(t). \end{aligned}$$

- $X_i(t)$ is a stochastic process associated with $\mathbb{P}^{(1)} \otimes \mathbb{P}^{(2)}$.
- We further define the random variables $\xi_{l,i}$ as follows

$$\xi_{l,i} \overset{\text{def}}{=} \frac{1}{\sqrt{2\lambda_l}} \cdot \int_0^T X_i(t)\phi_l(t)dt = \theta_l + \left(\frac{Z_{l,i}^{(1)} - Z_{l,i}^{(2)}}{\sqrt{2}}\right)$$

where $\theta_l = \frac{1}{\sqrt{2\lambda_l}} \int_0^T \left\{m_{\nu^*}^{(1)}(t) - m_{\nu^*}^{(2)}(t)\right\} \phi_l(t)dt.$

Hypothesis Testing

$$\xi_{l,i} \stackrel{\text{def}}{=} \frac{1}{\sqrt{2\lambda_l}} \cdot \int_0^T X_i(t)\phi_l(t)dt = \theta_l + \left(\frac{Z_{l,i}^{(1)} - Z_{l,i}^{(2)}}{\sqrt{2}}\right)$$

$$\text{where} \quad \theta_l = \frac{1}{\sqrt{2\lambda_l}} \int_0^T \left\{ m_{\nu^*}^{(1)}(t) - m_{\nu^*}^{(2)}(t) \right\} \phi_l(t)dt.$$

- $m_{\nu^*}^{(1)}(t) = m_{\nu^*}^{(2)}(t)$ for all $t \iff \theta_1 = \theta_2 = \cdots = 0$.
- $\xi_{l,i}$ are of mean θ_l and variance 1.
- $\xi_{1,i}, \xi_{2,i}, \ldots, \xi_{l,i}, \ldots$ are mutually uncorrelated (across index l).

Hypothesis Testing

- We test the following approximate hypothesis

$$\boxed{\widehat{H_0}: \quad \theta_1 = \theta_2 = \cdots = \theta_L = 0,}$$

$$\text{where } L \overset{\text{def}}{=} \min\left\{ l \in \mathbb{N} \,\middle|\, \frac{\sum_{l'=1}^{l} \lambda_{l'}}{\sum_{l''=1}^{\infty} \lambda_{l''}} > 0.95 \right\}.$$

- $\widehat{H_0} \iff \xi_{1,i}, \ldots, \xi_{L,i}$ are of mean zero.
- We implement the following asymptotic χ^2-test (Algorithm 1)

$$\boxed{\sum_{l=1}^{L} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_{l,i} \right)^2 > \chi^2_{L,1-\alpha}.}$$

- Highly-nonparametric hypothesis testing
  $\Longrightarrow$ normal distribution-based hypothesis testing.

Permutation Test

Recall our assumption[9] $\kappa_{\nu^*}^{(1)}(s,t) = \kappa_{\nu^*}^{(2)}(s,t) \stackrel{\text{def}}{=} \kappa(s,t)$.

- Violation (in the numerical sense) of the assumption will induce type-I error inflation.
- To reduce inflation, we need a permutation trick.
- Permutation test (Algorithm 2): We first shuffle (permute) the group labels $j \in \{1,2\}$ of shapes $\{K_i^{(j)}\}_{i=1}^n$; then, we apply Algorithm 1 to the permuted shapes.
- That is, Algorithm 2 = permutation + Algorithm 1.

[9] This assumption corresponds to the null hypothesis $\mathbb{P}^{(1)} = \mathbb{P}^{(2)}$.

Simulations

- For each $\varepsilon \in [0, 0.1]$, the distribution $\mathbb{P}^{(\varepsilon)}$ generates the following shapes

$$K_i^{(\varepsilon)} \stackrel{\text{def}}{=} \left\{ x \in \mathbb{R}^2 \,\middle|\, \inf_{y \in S_i^{(\varepsilon)}} \|x - y\| \leq \frac{1}{5} \right\}, \quad \text{where}$$

$$\begin{aligned} S_i^{(\varepsilon)} = & \left\{ \left(\frac{2}{5} + a_{1,i} \cdot \cos t, b_{1,i} \cdot \sin t \right) \,\middle|\, \frac{1-\varepsilon}{5}\pi \leq t \leq \frac{9+\varepsilon}{5}\pi \right\} \\ & \bigcup \left\{ \left(-\frac{2}{5} + a_{2,i} \cdot \cos t, b_{2,i} \cdot \sin t \right) \,\middle|\, \frac{6\pi}{5} \leq t \leq \frac{14\pi}{5} \right\}, \end{aligned}$$

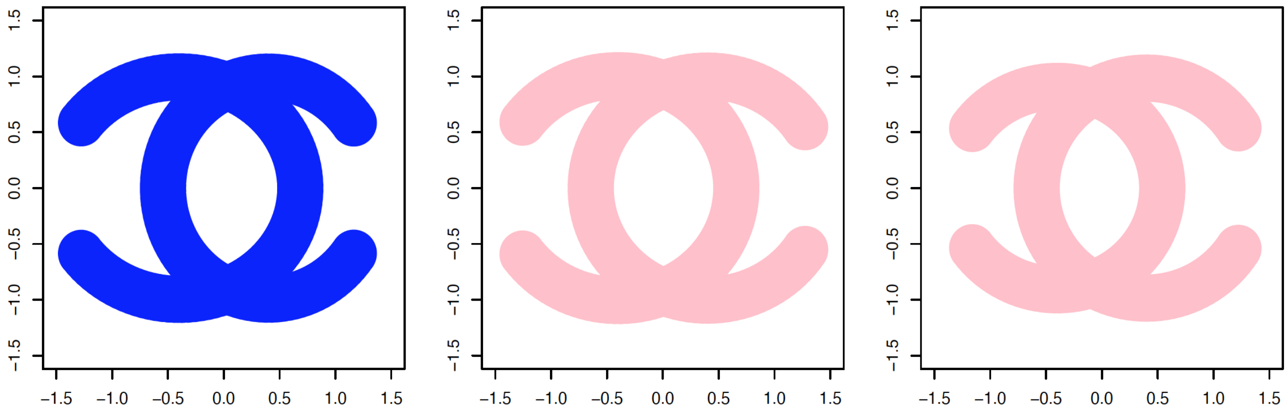
  where $a_{1,i},\, a_{2,i},\, b_{1,i},\, b_{2,i} \overset{i.i.d.}{\sim} N(0, 0.05^2)$.
- We test the following hypotheses

$$\begin{aligned} & H_0 : m_\nu^{(0)}(t) = m_\nu^{(\varepsilon)}(t) \text{ for all } (\nu, t) \in \mathbb{S}^{d-1} \times [0, T], \\ & vs. \;\; H_1 : m_\nu^{(0)}(t) \neq m_\nu^{(\varepsilon)}(t) \text{ for some } (\nu, t). \end{aligned}$$

$\varepsilon = 0.075$

$$100 \text{ blue shapes} \overset{iid}{\sim} \mathbb{P}^{(0)};$$
$$100 \text{ pink shapes} \overset{iid}{\sim} \mathbb{P}^{(\varepsilon)}.$$

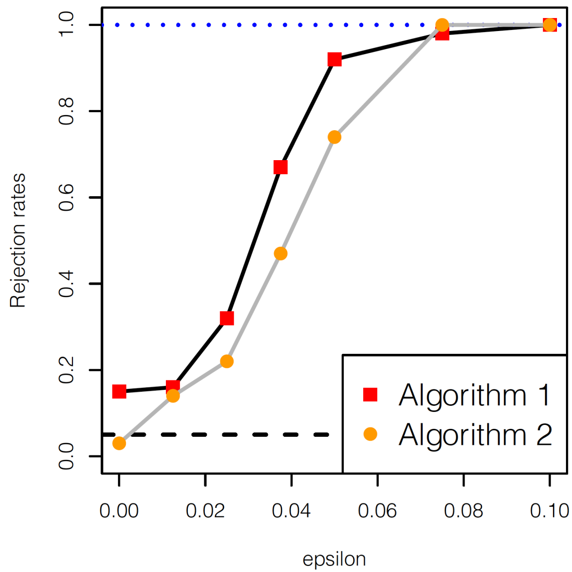
ε measures the discrepancy between the null hypothesis and the true shape-generating mechanism.



$$H_0 : m_\nu^{(0)}(t) = m_\nu^{(\varepsilon)}(t) \text{ for all } (\nu, t) \in \mathbb{S}^{d-1} \times [0, T],$$
$$vs. \quad H_1 : m_\nu^{(0)}(t) \neq m_\nu^{(\varepsilon)}(t) \text{ for some } (\nu, t),$$

Simulations ($\alpha = 0.05$)

$$H_0 : m_\nu^{(0)}(t) = m_\nu^{(\varepsilon)}(t) \text{ for all } (\nu, t) \in \mathbb{S}^{d-1} \times [0, T],$$
$$vs. \quad H_1 : m_\nu^{(0)}(t) \neq m_\nu^{(\varepsilon)}(t) \text{ for some } (\nu, t).$$



ε measures the discrepancy between the null hypothesis and the true shape-generating mechanism.

Section 3: Applications

Data Analysis I: Silhouette Database

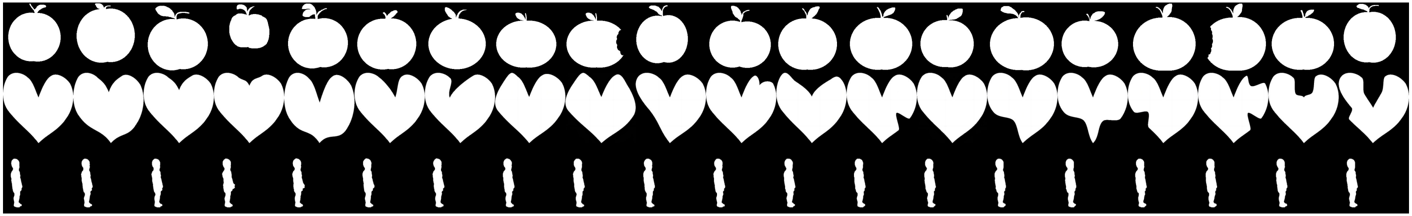


Table 1: P-values of Algorithms 1 and 2 for the silhouette database.

	Algorithm 1	Algorithm 2
Apples vs. Hearts	< 0.01	< 0.01
Apples vs. Children	< 0.01	< 0.01
Hearts vs. Children	< 0.01	< 0.01
Apples vs. Apples	0.26 (0.23)	0.46 (0.27)
Hearts vs. Hearts	0.17 (0.16)	0.47 (0.29)
Children vs. Children	0.39 (0.28)	0.49 (0.30)

Data Analysis II: Teeth of Primates

Data Analysis II: Teeth of Primates

Table 2: P-values of Algorithms 1 and 2 for the data set of mandibular molars.

	Algorithm 1	Algorithm 2
Tarsius vs. Microcebus	$< 10^{-3}$	$< 10^{-3}$
Tarsius vs. Mirza	$< 10^{-3}$	$< 10^{-3}$
Tarsius vs. Saimiri	$< 10^{-3}$	$< 10^{-3}$
Microcebus vs. Mirza	$< 10^{-3}$	0.009
Microcebus vs. Saimiri	$< 10^{-3}$	$< 10^{-3}$
Mirza vs. Saimiri	$< 10^{-3}$	$< 10^{-3}$
Tarsius vs. Tarsius	0.206 (0.195)	0.519 (0.274)

Section 4: Mathematical Foundations

Polish Space-valued Random Variables

- Ω := the collection of shapes in $\mathbb{R}^d$ satisfying some topological conditions[10]. They are the shapes of interest.
- For each $K \in \Omega$ and fixed direction $\nu \in \mathbb{S}^{d-1}$, we proved that $t \mapsto \mathrm{SECT}(K)(\nu, t)$ belongs to the Sobolev space $\mathcal{H} := H_0^1(0,T)$, i.e., $\mathrm{SECT}(K)(\nu, \cdot) \in \mathcal{H}$.
- (Sobolev spaces are usually implemented to show the well-posedness of PDEs.[11] By Sobolev embedding theorm, $\mathcal{H} = H_0^1(0,T)$ is a RKHS)
- In addition, we proved the continuity[12] of the following function

$$\mathbb{S}^{d-1} \to \mathcal{H}, \quad \nu \mapsto \mathrm{SECT}\, K(\nu, \cdot),$$

that is, $\mathrm{SECT}(K) \in C(\mathbb{S}^{d-1}; \mathcal{H})$.

[10] They involve too much machinery of computational topology, hence, are omitted.

[11] e.g., Junfeng Li and **Kun Meng**. Global well-posedness for the fifth order Kadomtsev-Petviashvili II equation in three-dimensional space. *Nonlinear Analysis*, 130: 157-175, 2016.

[12] Precisely, 1/2-Hölder continuity

Polish Space-valued Random Variables

$$\begin{aligned}\mathrm{SECT}: \ \Omega &\to C(\mathbb{S}^{d-1};\mathcal{H}),\\ K &\mapsto \mathrm{SECT}(K)\end{aligned}$$

- Hence, SECT(K) takes values in $C(\mathbb{S}^{d-1};\mathcal{H})$, which is a separable Banach space (hence, Polish space, suitable for probability).
- We defined a metric (not a semi-metric) on Ω as follows

$$\rho(K_1,K_2) \overset{\text{def}}{=} \sup_{\nu\in\mathbb{S}^{d-1}} \left\{ \left(\int_0^T \left| \chi(K^{\nu}_{1,\tau}) - \chi(K^{\nu}_{2,\tau}) \right|^2 d\tau \right)^{1/2} \right\}.$$

- The map $\mathrm{SECT}: \Omega \to C(\mathbb{S}^{d-1};\mathcal{H})$ is Borel-measurable, hence, a random variable.
- The conditions of the Karhunen–Loève expansion are satisfied.

Conclusions

- **Methodology**: We proposed statistical inference methods for testing whether two collections of shapes are significantly different.
- Our discussions connect the following fields: algebraic and computational topology, probability theory and stochastic processes, Sobolev spaces and functional analysis, statistical inference, and morphology.
- Our results have been posted on arXiv.[13]
- **Future work**: We will apply similar approaches to grayscale images of tumors[14] and fMRI data[15].

[13] **Kun Meng**, Jinyu Wang, Lorin Crawford, and Ani Eloyan. Randomness and statistical inference of shapes via the smooth Euler characteristic transform. arXiv preprint arXiv: 2204.12699 (2023). (Submitted to JASA — major revisions)

[14] Q. Jiang, S. Kurtek, and T. Needham. The weighted euler curve transform for shape and image analysis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pages 844–845, 2020.

[15] **Kun Meng** and Ani Eloyan. Population-level task-evoked functional connectivity via Fourier analysis. arXiv preprint arXiv: 2102.12039 (2022).

Thank You !